# ETSI TR 104 159 V1.1.1 (2026-01)

TECHNICAL REPORT

# Securing Artificial Intelligence (SAI); Understanding and Preventing Harm from Generative AI

Reference
DTR/SAI-0019

Keywords
AI, cybersecurity, end-user, privacy

***ETSI***

650 Route des Lucioles
F-06921 Sophia Antipolis Cedex - FRANCE

Tel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16

Siret N° 348 623 562 00017 - APE 7112B
Association à but non lucratif enregistrée à la
Sous-Préfecture de Grasse (06) N° w061004871

# Contents

# Intellectual Property Rights

## Essential patents

IPRs essential or potentially essential to normative deliverables may have been declared to ETSI. The declarations pertaining to these essential IPRs, if any, are publicly available for **ETSI members and non-members**, and can be found in ETSI SR 000 314: *"Intellectual Property Rights (IPRs); Essential, or potentially Essential, IPRs notified to ETSI in respect of ETSI standards"*, which is available from the ETSI Secretariat. Latest updates are available on the ETSI IPR online database.

Pursuant to the ETSI Directives including the ETSI IPR Policy, no investigation regarding the essentiality of IPRs, including IPR searches, has been carried out by ETSI. No guarantee can be given as to the existence of other IPRs not referenced in ETSI SR 000 314 (or the updates on the ETSI Web server) which are, or may be, or may become, essential to the present document.

## Trademarks

The present document may include trademarks and/or tradenames which are asserted and/or registered by their owners. ETSI claims no ownership of these except for any which are indicated as being the property of ETSI, and conveys no right to use or reproduce any trademark and/or tradename. Mention of those trademarks in the present document does not constitute an endorsement by ETSI of products, services or organizations associated with those trademarks.

**DECT™**, **PLUGTESTS™**, **UMTS™** and the ETSI logo are trademarks of ETSI registered for the benefit of its Members. **3GPP™**, **LTE™** and **5G™** logo are trademarks of ETSI registered for the benefit of its Members and of the 3GPP Organizational Partners. **oneM2M™** logo is a trademark of ETSI registered for the benefit of its Members and of the oneM2M Partners. **GSM®** and the GSM logo are trademarks registered and owned by the GSM Association.

# Foreword

This Technical Report (TR) has been produced by ETSI Technical Committee Securing Artificial Intelligence (SAI) (SRdAP).

# Modal verbs terminology

In the present document "**should**", "**should not**", "**may**", "**need not**", "**will**", "**will not**", "**can**" and "**cannot**" are to be interpreted as described in clause 3.2 of the ETSI Drafting Rules (Verbal forms for the expression of provisions).

"**must**" and "**must not**" are **NOT** allowed in ETSI deliverables except when used in direct citation.

# 1 Scope

The present document provides an understanding of the harm from Generative AI, along with presenting the different ways to prevent that harm. This includes but is not limited to malicious code generation, deepfakes, spam messages, disinformation, etc. The areas also covered are the issues of AI hallucinations, loss of confidentiality and IPR infringements. The types of methods to counter the harm from Generative AI to be included are detection, enforcement, reporting and removal.

# 2 References

## 2.1 Normative references

Normative references are not applicable in the present document.

## 2.2 Informative references

References are either specific (identified by date of publication and/or edition number or version number) or non-specific. For specific references, only the cited version applies. For non-specific references, the latest version of the referenced document (including any amendments) applies.

NOTE: While any hyperlinks included in this clause were valid at the time of publication, ETSI cannot guarantee their long-term validity.

The following referenced documents may be useful in implementing an ETSI deliverable or add to the reader's understanding, but are not required for conformance to the present document.

[i.1] What is Gen AI? Generative AI explained.

[i.2] Voluntary AI Safety Standard: The 10 guardrails.

[i.3] Bill No. 2338/2023: "Regulatory framework for artificial intelligence passes in Brazil's Senate".

[i.4] Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems.

[i.5] AI Watch: Global regulatory tracker – China.

[i.6] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance).

[i.7] RESPONSIBLE AI #AIFORALL; Approach Document for India Part 1 – Principles for Responsible AI.

[i.8] Hiroshima Process International Guiding Principles for Organizations Developing Advanced AI System.

[i.9] AI Guidelines for Business Ver1.0; April 19, 2024; Ministry of Internal Affairs and Communications Ministry of Economy, Trade and Industry.

[i.10] A New Era for AI: Republic of Korea Takes a Bold Step with AI Regulation.

[i.11] Code of Practice for the Cyber Security of AI; 2025; UK GOV Department for Science, Innovation & Technology.

[i.12] California's AB 2013: "Generative artificial intelligence: training data transparency".

[i.13] SB-942 California AI Transparency Act.

[i.14] AB-2602 Contracts against public policy: personal or professional services: digital replicas.

[i.15] Colorado's Landmark AI Act: What Companies Need To Know.

[i.16] Tennessee Law Addresses Proliferation of Deepfakes.

[i.17] Utah Enacts AI-Focused Consumer Protection Bill.

[i.18] Generative AI Navigating Intellectual Property.

[i.19] Artificial intelligence and copyright: use of generative AI tools to develop new content; European Innovation Council and SMEs Executive Agency.

[i.20] ETSI TR 104 048 (V1.1.1): "Securing Artificial Intelligence (SAI); Data Supply Chain Security".

[i.21] What are AI hallucinations?

[i.22] Data poisoning: how artists are sabotaging AI to take revenge on image generators.

[i.23] Indirect Prompt Injection: Generative AI's Greatest Security Flaw; Matt Sutton, Damian Ruck; 2024; Centre for Emerging Technology and Security at The Alan Turing Institute.

[i.24] Security Guidelines for Generative Artificial Intelligence Application Service; ITU-T SG17.

[i.25] Implementation guidelines for digital watermarking; ITU-T SG17.

[i.26] Notice on Issuing the Measures for Identifying Synthetic Content Generated by Artificial Intelligence.

[i.27] Cybersecurity technology — Labelling method for content generated by artificial intelligence; 2025; tc260.

[i.28] AB-1836 Use of likeness: digital replica.

[i.29] ETSI TS 102 165-1 (V5.3.1): "Cyber Security (CYBER); Methods and protocols; Part 1: Method and pro forma for Threat, Vulnerability, Risk Analysis (TVRA)".

[i.30] ETSI TS 104 102: "Cyber Security (CYBER); Encrypted Traffic Integration (ETI); ZT-Kipling methodology".

[i.31] Tackling deepfakes in European policy; 2021; European Parliamentary Research Service.

[i.32] Increasing Threat of Deepfake Identities.

[i.33] GenAI and the battle against misinformation; 2024; Yash Shreshtha; Duke Corporate Education.

[i.34] LLM09:2025 Misinformation.

[i.35] AI and GDPR: the CNIL publishes new recommendations to support responsible innovation.

[i.36] Adversarial Misuse of Generative AI; 2025; Google Threat Intelligence Group.

[i.37] Evaluating Malicious Generative AI Capabilities; 2024 Centre for Emerging Technology and Security; The Alan Turing Institute.

[i.38] Hackers exploit generative AI; 2024; Centre for Cyber Security.

[i.39] ETSI TS 104 119: "Methods for Testing & Specification (MTS); AI Testing Guidelines for Documentation of AI-enabled Systems".

[i.40] Red Teaming for GenAI Harms; 2024; Ofcom.

[i.41] Data Authenticity, Consent, and Provenance for AI Are All Broken: What Will It Take to Fix Them?

[i.42] Deepfake Defences; 2024; Ofcom.

[i.43] ETSI TS 102 232 (all parts): "Lawful Interception (LI); Handover Interface and Service-Specific Details (SSD) for IP delivery".

[i.44] ETSI TS 104 223: "Securing Artificial Intelligence (SAI); Baseline Cyber Security Requirements for AI Models and Systems".

[i.45] ETSI TR 104 128: "Securing Artificial Intelligence (SAI); Guide to Cyber Security for AI Models and Systems".

[i.46] ETSI EN 304 223: "Securing Artificial Intelligence (SAI); Baseline Cyber Security Requirements for AI Models and Systems".

[i.47] Tennessee Personal Rights Protection Act of 1984.

[i.48] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

# 3 Definition of terms, symbols and abbreviations

## 3.1 Terms

For the purposes of the present document, the following terms apply:

**agentic AI:** small, specialized pieces of software that can make decisions and operate cooperatively or independently to achieve system objectives

NOTE: Agentic AI refers to AI systems composed of agents that can behave and interact autonomously to achieve their objectives.

**confidentiality:** preserving authorized restrictions on access and disclosure, including means for protecting personal privacy and proprietary information

**copyright:** protection for original works of authorship as soon as an author fixes the work in a tangible form of expression

**detection:** fact of noticing or discovering something

**generative artificial intelligence:** deep-learning models that can generate high-quality text, images, and other content based on the data they were trained on

**harm:** to hurt someone or damage something

**intellectual property rights:** any and all rights associated with intangible assets owned by a person or company and protected against use without consent

**legislation:** rules or laws relating to a particular activity that are made by a government

**malicious:** intent to cause harm or damage

**misinformation:** wrong information or information intended to deceive

**open-source model:** binaries of machine learning algorithms pre-trained on often-large datasets to achieve state-of-the-art performance in a machine learning application that are released to the public for everyone to use, for either model inference or transfer learning

**prevention:** act of stopping something from happening or of stopping someone from doing something

**regulation:**rule or directive made and maintained by an authority

**spam:** unwanted email or messages

## 3.2 Symbols

Void.

## 3.3 Abbreviations

For the purposes of the present document, the following abbreviations apply:

| | |
|---|---|
| AI | Artificial Intelligence |
| AB | Assembly Bill |
| C2PA | Coalition for Content Provenance and Authenticity |
| CAIA | Colorado Artificial Intelligence Act |
| CNIL | Commission Nationale de l'Informatique et des Libertés |
| CSAM | Child Sexual Abuse Material |
| DKIM | DomainKeys Identified Mail |
| DMARC | Domain-based Message Authentication Reporting and Conformance |
| DMA | Diffusion Model Architecture |
| ELVIS | Ensuring Likeness Image and Voice Security |
| EU | European Union |
| FBI | Federal Bureau of Investigation |
| GAN | Generative Adversarial Network |
| GDPR | General Data Protection Regulation |
| GenAI | Generative Artificial Intelligence |
| HAIP | Hiroshima AI Process |
| HIC | Human Interaction Component |
| HITL | Human in the Loop |
| HOTL | Human on the Loop |
| HTTPS | Hypertext Transfer Protocol Secure |
| IC3 | Internet Crime Complaint Centre |
| ICT | Information and Communications Technology |
| IP | Intellectual Property |
| IP | Internet Protocol |
| ISCC | International Standard Content Code |
| ISO | International Organization for Standardization |
| LMM | Large Language Models |
| MFA | Multi-Factor Authentication |
| NCII | Non-Consensual Intimate Image |
| NLP | Natural Language Processing |
| NSFW | Not Safe For Work |
| PET | Parameter-Efficient Tuning |
| RAG | Retrieval-Augmented Generation |
| SB | Senate Bill |
| SSD | Service Specific Details |
| S/MIME | Secure/Multipurpose Internet Mail Extensions |
| SMTPS | Simple Mail Transfer Protocol Secure |
| SPF | Sender Policy Framework |
| SSL | Secures Sockets Layer |
| TLS | Transport Layer Security |
| TVRA | Threat, Vulnerability, Risk Analysis |
| UK | United Kingdom |
| US/USA | United States of America |
| ZT | Zero Trust |

# 4 Introduction

## 4.1 What is Generative Artificial Intelligence (GenAI)

Generative Artificial Intelligence (generative AI, GenAI, or GAI) is a subset of artificial intelligence that can use Generative Adversarial Models (GANs) or Diffusion Model Architecture (DMAs) to produce text, images, videos, or other forms of data [i.1]. Instead of being based on the input, these models learn the underlying patterns and structures of their training data and use them to produce new data, instead of being based on the input, often in the form of natural language prompts. In general, Gen AI uses algorithms to organize large, complex data sets into potentially meaningful clusters of information to create new content, including text, images and audio, in response to a query or prompt.

NOTE: The issues of harm, the threats and mitigations that apply to GenAI are also relevant to Agentic AI as Agentic AI is an evolution of GenAI.

Related work is also under development in ITU-T SG17 [i.25] on 'Security guidelines for Generative Artificial Intelligence Application Service' [i.24] and 'Implementation guidelines for digital watermarking', which complement the subjects discussed in the present document.

## 4.2 Uses of GenAI

Generative AI is used across various industries, including - but not limited to - software development, healthcare, finance, entertainment, customer service, sales and marketing, art, writing, fashion, and product design. Some examples of use cases of GenAI are:

1) Text: capable of natural language processing, machine translation, and natural language generation and can be used as a foundation model for other tasks.

2) Code: Generate source code for programs.

3) Images: Commonly used for text-to-image generation and neural style transfer.

4) Audio: produces natural-sounding speech synthesis and text-to-speech capabilities.

5) Video: Can generate temporally coherent, detailed and photorealistic video clips.

## 4.3 Impact of Regulation and Legislation in a Global Perspective

### 4.3.1 Introduction

Due to the rapid development of generative AI in a short period of time, numerous regulations and legislation have been passed to prevent harm from AI and ensure responsible use and development. Some of these are broad measures while others have a narrower focus. This clause will highlight their impact on generative AI and the measures developers may have to take to be compliant. There is an overlap between these different measures which means if an organization is compliant with one it could be compliant or partially compliant with another. The following clauses are a non-exhaustive list and represent a snapshot of existing regulations and legislation at the time of publication.

### 4.3.2 Australia - Voluntary AI Safety Standard

The Voluntary AI Safety Standard [i.2] gives practical guidance to all Australian organizations on how to safely and responsibly use and innovate with Artificial Intelligence (AI). The standard consists of 10 voluntary guardrails that apply to all organizations throughout the AI supply chain. They include transparency and accountability requirements across the supply chain. They also explain what developers and deployers of AI systems need to do. The guardrails are to help organizations benefit from AI while mitigating and managing the risks that AI may pose to organizations, people and groups.

An example guardrail from the Voluntary AI Safety Standard [i.2]: "*Inform end-users regarding AI-enabled decisions, interactions with AI and AI-generated content. - Create trust with users. Give people, society and other organizations confidence that you are using AI safely and responsibly. Disclose when you use AI, its role and when you are generating content using AI. Disclosure can occur in many ways. It is up to the organization to identify the most appropriate mechanism based on the use case, stakeholders and technology used*".

### 4.3.3 Brazilian Legal Framework for Artificial Intelligence, Marco Legal da Inteligência Artificial (Bill No. 2338/2023)

Bill No. 2338/2023 [i.3] to establish a national regulatory framework covering the development, use, and governance of AI systems in Brazil. The text reflects a commitment to the centrality of the human person, responsible innovation, AI market competitiveness, and the implementation of safe and reliable systems. The regulatory framework defines a set of rights designed to protect individuals or groups affected by AI systems, including generative AI, such as:

- The right to clear, accessible information about the use of AI in their interactions with such systems.
- The right to request reviews of automated decisions by humans in certain circumstances.
- The right to non-discrimination (illicit or abusive), as well as the right to have direct or indirect discriminatory bias corrected.

### 4.3.4 Canada - Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems

The code [i.4] identifies measures that should be applied by all organizations developing or managing the operations of a generative AI system with general-purpose capabilities, as well as additional measures that should be taken by organizations developing or managing the operations of these systems that are made widely available for use, and which are therefore subject to a wider range of potentially harmful or inappropriate use. Organizations developing and managing the operations of these systems both have important and complementary roles. Developers and managers need to share relevant information to ensure that adverse impacts can be addressed by the appropriate actor.

In undertaking this voluntary commitment, developers and managers of advanced generative systems commit to working to achieve the following outcomes:

1) Accountability - Organizations understand their role with regard to the systems they develop or manage, put in place appropriate risk management systems, and share information with other organizations as needed to avoid gaps.
2) Safety - Systems are subject to risk assessments, and mitigations needed to ensure safe operation are put in place before deployment.
3) Fairness and Equity - Potential impacts concerning fairness and equity are assessed and addressed at different phases of the development and deployment of the systems.
4) Transparency - Sufficient information is published to allow consumers to make informed decisions and for experts to evaluate whether risks have been adequately addressed.
5) Human Oversight and Monitoring - System use is monitored after deployment, and updates are implemented as needed to address any risks that materialize.
6) Validity and Robustness - Systems operate as intended, are secure against cyber-attacks, and their behaviour in response to the range of tasks or situations to which they are likely to be exposed is understood.

### 4.3.5 China

#### 4.3.5.1 The Interim Measures for the Management of Generative AI Services

Under the AI Measures, "generative AI technology" [i.5] refers to models and related technology that have the ability to generate text, images, audio, videos, or other content. The key roles under the AI Measures are Generative AI service providers and users. "Generative AI service provider" refers to any organization or individual that utilizes generative AI technology to provide generative AI services (including providing such services through the provision of a programmable interface or other means).

"User of Generative AI services" refers to any organization or individual that uses Generative AI services to generate content. The AI Measures do not apply if an organization or institution engages in research, development, or application of generative AI technology but does not offer Generative AI services to the domestic public in China.

Examples of the key compliance requirements:

1) Content moderation: Generative AI service providers are required to promptly remove any illegal content, employ measures for model optimization training, and report cases to the relevant authorities.

2) Reporting mechanism: Generative AI service providers need to establish a complaints and reporting mechanism, where they accept and handle complaints and reports from the public and provide feedback on the outcome of these cases.

#### 4.3.5.2 Measures for the Labelling of Artificial Intelligence-Generated and Synthetic Content

The Measures standardize requirements for providers of generation and synthesis services to add explicit and implicit labels (as applicable) to generated synthetic content, including texts, images, audio, videos and virtual scenes. The use of explicit labels (which are clearly visible to users) and implicit labels (which are embedded in the content's metadata) in the Measures [i.26].

#### 4.3.5.3 GB 45438-2025 Cybersecurity Technology - Labelling Method for Content Generated by Artificial Intelligence

This reference [i.27] is a complement to the Measures [i.5] (see clause 4.3.5.2) as a mandatory standard. It specifies the format of explicit labels required by the measures, such as inserting "AI" by text, superscript, voice and rhythm, as well as the metadata to be added as implicit labels.

### 4.3.6 EU: Artificial Intelligence Act

The AI Act (Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence) [i.6] is a legal framework for AI worldwide. The rules aim to foster trustworthy AI in Europe. The AI Act sets out a clear set of risk-based rules for AI developers and deployers regarding specific uses of AI.

It includes requirements to disclose copyrighted material used to train generative AI systems and to label any AI-generated output as such.

NOTE: Before the AI Act was passed, several EU countries produced their own regulation on AI. These have been deprecated in favour of the AI Act upon its coming into force.

### 4.3.7 India NITI Aayog: Part 1 Principles for Responsible AI

It identifies the following broad principles for responsible management of AI [i.7]:

1) Principle of Safety and Reliability.

2) Principle of Equality.

3) Principle of Inclusivity and Non-discrimination.

4) Principle of Privacy and Security.

5) Principle of Transparency.

6) Principle of Accountability.

7) Principle of protection and reinforcement of positive human values.

## 4.3.8 Japan

### 4.3.8.1 Hiroshima AI Process: International Guiding Principles for Organizations Developing Advanced AI Systems

The Hiroshima AI Process (HAIP) is a set of guidelines and principles for developing and using AI systems safely and responsibly.

The Hiroshima Process International Guiding Principles for Organizations Developing Advanced AI Systems [i.8] aims to promote safe, secure, and trustworthy AI worldwide and provides guidance for organizations developing and using the most advanced AI systems, including the most advanced foundation models and generative AI systems. Organizations may include, among others, entities from academia, civil society, the private sector, and the public sector.

Examples of the principles include:

- Develop and deploy reliable content authentication and provenance mechanisms, where technically feasible, such as watermarking or other techniques to enable users to identify AI-generated content.
- Prioritize research to mitigate societal, safety and security risks and prioritize investment in effective mitigation measures.
- Implement appropriate data input measures and protections for personal data and intellectual property.

### 4.3.8.2 AI Guidelines for Business Version 1.0

The Guidelines [i.9] present unified guiding principles in AI governance in Japan to promote the safe and secure use of AI. It is intended to help people who use AI in various businesses to fully recognize AI risks based on international trends and stakeholders' concerns and to voluntarily take the necessary countermeasures across the entire lifecycle.

One of the key guiding principles is the Human-Centric use of AI. This includes when developing, providing, or using an AI system or service, each AI business actor should act in a way that does not violate the human rights guaranteed by the Constitution of Japan or granted internationally:

1) Respect human dignity and the autonomy of individuals.

2) Paying attention to manipulations by AI on decision-making and emotions.

3) Countermeasures against disinformation, misinformation, and biased information generated by AI.

4) Ensuring diversity/inclusion for example adopting universal design, ensuring accessibility, and providing relevant stakeholders with education and support.

5) Providing user support.

6) Ensuring sustainability.

## 4.3.9 South Korea - Framework Act on Artificial Intelligence Development and Establishment of a Foundation for Trustworthiness (AI Framework Act)

The act [i.10] aims to protect citizens' rights and dignity, improve their quality of life, and strengthen national competitiveness by regulating fundamental matters necessary for the sound development of AI and the establishment of a foundation of trust.

It defines AI as "the electronic implementation of human intellectual abilities such as learning, reasoning, perception, judgment, and language understanding" and AI systems as "AI-based systems that infer outputs such as predictions, recommendations, and decisions that affect real and virtual environments for given objectives, with varying levels of autonomy and adaptability".

Generative AI is one of the key provisions of this act. Which is defined as systems producing text, images, videos, or other outputs based on the structure and characteristics of the input data. It specifies requirements for AI safety and trustworthiness for generative AI as when providing products or services utilizing generative AI, AI businesses need to notify users in advance of such fact. AI businesses also need to label the outputs of such products or services clearly as AI-generated, particularly when the outputs mimic real-world sounds, images, or videos. For artistic or creative expressions, this obligation can be fulfilled in a manner that does not interfere with the display or appreciation of the work.

## 4.3.10 UK: AI Code of Practice

The scope of this voluntary Code of Practice [i.11] is focused on AI systems. This includes systems that incorporate deep neural networks, such as generative AI. The Code sets out cybersecurity requirements for the lifecycle of AI. These are secure design, secure development, secure deployment, secure maintenance and secure end of life.

It has been developed into ETSI TS 104 223 [i.44] Baseline Cyber Security Requirements for AI Models and Systems. This TS establishes baseline cybersecurity requirements for AI models and systems that enable them to embed cybersecurity and resilience across the AI lifecycle. This TS is supported by ETSI TR 104 128 [i.45], which provides an implementation guide for organizations implementing baseline cybersecurity requirements for AI. ETSI TS 104 223 [i.44], in turn, has been transposed into ETSI EN 304 223 [i.46].

## 4.3.11 USA

### 4.3.11.1 California

#### 4.3.11.1.1 AB 2013 Generative artificial intelligence: training data transparency

AB 2013 [i.12] requires developers of GenAI systems to publicly disclose detailed information about the datasets used in their development on their website. The law's transparency mandate applies to all GenAI systems and services made available to Californians, regardless of whether compensation is involved, provided the systems were released on or after January 1, 2022. Compliance becomes mandatory by January 1, 2026, with updates required for substantial modifications to existing systems.

The statute defines GenAI broadly, encompassing systems that generate synthetic content, such as text, images, or audio, modelled after training data. Documentation to be posted on the developer's website need to include (among other things):

- A high-level summary of datasets used in the system's development.
- Information on dataset sources, ownership, and intended purpose.
- Descriptions of data types, including whether they include data protected by copyright, trademark, or patents, and whether the data includes personal information or aggregated consumer data.
- Details on synthetic data usage, dataset cleaning or processing, and whether datasets were purchased or licensed.
- The timeframes for data collection and the date datasets were first used.
- Whether synthetic data generation was used in developing the GenAI system or the service.

#### 4.3.11.1.2 SB 942 California AI Transparency Act

The California AI Transparency Act ("SB 942") [i.13] creates transparency mechanisms that allow consumers to determine whether an "image, video, or audio content or content that is any combination thereof, was created or altered" using generative artificial intelligence.

It imposes three core requirements:

1) The Covered Provider need to make available a free, publicly available GenAI content detection tool that allows users "to assess whether image, video, or audio content or content that is any combination thereof, was created or altered by" the Covered Provider's GenAI system. The tool needs to allow users to upload content or provide a URL to the content needing detection. It also requires the Covered Provider to collect user feedback regarding the tool's efficacy to improve it.

2) The Covered Provider need to give users the option to include:

   i) a non-hidden (i.e. manifest) disclosure for a GenAI image, video, or audio content that the user generates; and

   ii) a hidden (i.e. latent) disclosure for any user-generated GenAI image, video, or audio content.

   Such hidden disclosure needs to be detectable with the tool under the first requirement.

3) The Covered Provider need to contractually require any third-party licensees of the Covered Provider's GenAI system to maintain the second requirement's disclosure capabilities. If a Covered Provider knows that a third-party licensee has removed the disclosure capabilities from the GenAI system, the Covered Provider need to revoke the third-party's license within 96 hours.

#### 4.3.11.1.3 AB 1836 and AB 2602

Assembly Bill 2602 (AB 2602) [i.14] and Assembly Bill 1836 (AB 1836) [i.28], establish fresh regulatory requirements focusing on transparency, accountability, and ethical AI use, particularly in the entertainment industry. AB 2602 prevents the unauthorized use of digital replicas of individuals' voices or likenesses in contracts for personal or professional work, requiring specific consent and representation during negotiations. AB 1836 prohibits the creation or distribution of digital replicas of deceased personalities without permission from their estate, aiming to protect the posthumous rights of publicity.

### 4.3.11.2 Colorado - Colorado Artificial Intelligence Act (CAIA)

The CAIA [i.15] is primarily focused on high-risk artificial intelligence systems, which are defined as any system that, when deployed, makes or is a substantial factor in making a "consequential decision". The consequential decisions generally relate to those involving education, employment, financial services, housing, health care or legal services. The CAIA is designed to protect against algorithmic discrimination, namely, unlawful differential treatment that disfavours an individual or group based on protected characteristics. The law imposes various obligations relating to documentation, disclosures, risk analysis and mitigation, governance, and impact assessments for developers and deployers of high-risk AI systems. Concerning all AI systems that interact with consumers, deployers need to ensure that consumers are aware they are interacting with an AI system.

### 4.3.11.3 Tennessee - Ensuring Likeness Image and Voice Security (ELVIS) Act

Tennessee's Ensuring Likeness, Voice and Image Security (ELVIS) Act [i.16] aims to protect individuals from the use of their persona in connection with "deepfakes" (i.e. fake content generated by artificial intelligence (AI) that a user is likely to mistakenly believe is legitimate). It specifically addresses the deepfake issue in three ways.

First, it expands the state's existing personal rights law (the Personal Rights Protection Act of 1984 or PRRA [i.47]), which previously only protected a person's name, image and likeness, to explicitly include protections for an individual's "voice" (defined as an individual's actual voice as well as simulations of that voice). The act also expands PRRA by prohibiting any unauthorized "commercial use" of a person's personal rights; PRRA had previously been limited to uses "in advertising."

Second, the act creates a private right of action against anyone who unlawfully publishes, performs, distributes, transmits or otherwise makes available to the public an individual's voice or likeness, with the knowledge that the use of the voice or likeness was not authorized by the individual.

Third, it creates a private right of action against anyone who "*distributes, transmits, or otherwise makes available an algorithm, software, tool, or other technology, service, or device, the primary purpose or function of which is the production of an individual's photograph, voice, or likeness without authorization from the individual*." Where the individual is a minor, authorization is required from a parent or guardian, and where the individual is deceased, authorization is required from an executor or heir. While this provision does not mention AI explicitly, the focus of the prohibition can be assumed to include AI.

#### 4.3.11.4 Utah - Artificial Intelligence Policy Act

The bill imposes on companies operating in Utah including disclosure requirements on entities using generative artificial intelligence tools with their customers and limits an entity's ability to blame generative AI for statements that violate consumer protection laws [i.17].

# 5 Impact of GenAI on Intellectual Property Rights

## 5.1 Overview

Generally, Generative AI systems are typically trained on large available datasets which may include copyrighted works, personal information, biometric data, and harmful and illegal content [i.18]. AI developers have argued that such training is protected under fair use, while copyright holders have argued that it infringes their rights. There is ongoing litigation in various countries worldwide over whether the scraping, downloading and processing of materials, the trained AI models and their output involve breaches of IP, privacy and contract.

## 5.2 Copyright theft and infringement

The issues and problems of copyright are that AI generative software cannot create "original" media from zero; rather it needs a pre-existing training data input to bootstrap the algorithms through a machine learning model [i.19]. The origin of those images/input is crucial not only regarding the ownership of the copyrights generated but also to avoid potential copyright infringement of pre-existing works. In the case that the algorithm is trained with protected works owned by a third party (and protected via copyright, image rights or data protection) without authorization or without licensing its content, the results may infringe on the rights of these third parties.

Proponents of fair use training have argued that it is a transformative use and does not involve making copies of copyrighted works available to the public. Critics have argued that image generators can create nearly identical copies of some copyrighted images, and that generative AI programs compete with the content they are trained on.

A separate question is whether AI-generated works can qualify for copyright protection. The United States Copyright Office has ruled that works created by artificial intelligence without any human input cannot be copyrighted because they lack human authorship. However, the office has also begun taking public input to determine if these rules need to be refined for generative AI.

Regarding the ownership of the creations made with the use of generative AI tools, the answer will likely vary depending on the following aspects:

- The laws of the relevant jurisdiction that govern the AI and the creation of the work, if any.
- The extent of the role performed by both the human user and the AI platform in generating the output, as having an AI tool fully develop a new work is different from using such a tool to review a work or make slight adaptations to it; and
- the IP provisions under the Terms and conditions of the license with the service provider.

As a result of those variations, copyright over works developed with generative AI may belong:

- To the creators of the algorithm, who retain ownership over the works created by their algorithm.
- To the user of the AI tool (which is the most common).

- To no one (the works created through the generative AI tool are either considered not to be protected by copyright or need to be put in the public domain as per the terms and conditions of the AI tool used).

Several measures can be taken to check compliance at least in the EU with the AI Act and Directive 2019/790/EU on copyright in the Digital Single Market when using GenAI:

- Does the software provider confirm that the data used to train the algorithm has been legally accessed or licensed?
- Has the software provider been involved in any known copyright lawsuit?
- Is the AI capable of adapting or being trained with the assets, and works of the user?
- Will the platform own any rights over the creations?
- Who has the commercial exploitation rights over the results?
- Is there any disclaimer about infringement liabilities (i.e. does the AI tool provider exclude any liability for infringement of third-party copyright through the use of its tool)?
- Are there any close visual hits, after using a reverse image or text search from the results obtained?

## 5.3 Understanding the Training Material

### 5.3.1 Data curation

#### 5.3.1.1 Overview

In general, data curation is the ongoing processing and maintenance of data throughout its lifecycle to ensure long-term accessibility, sharing, and preservation [i.20]. With AI the data curation, or processing, stage typically includes a number of aggregation and transformation steps, including data storage, pre-processing, cleaning, enrichment and labelling. It can include integrating data from multiple sources and formats, identifying missing components of the data, removing errors and sources of noise, conversion of data into new formats, labelling the data, data augmentation using real and synthetic data, or scaling the data set using data synthesis approaches. When collecting the data for the training model is not just a security issue but also a data integrity issue. The techniques for assessing and understanding data quality for performance, transparency or ethics purposes apply to ensuring security assurance [i.20]. An example method for identifying threats and risks to the data supply chain is the Threat, Vulnerability, Risk Analysis (TVRA) [i.29] which defines a method primarily for use in undertaking an analysis of the threats, risks and vulnerabilities of an Information and Communications Technology (ICT) system to identify applicable countermeasures. This can also be supported by the ZT-Kipling method [i.30] which adopts a Zero Trust (ZT) within a business to promote transparent and explicable security provisions within that business.

#### 5.3.1.2 Cleaning and filtering

Data cleaning is the process of preparing data for the training model by removing or modifying data that is incorrect, incomplete, irrelevant, duplicated, or improperly formatted. Some general steps should be taken when preparing the data:

1) Remove duplicates.
2) Filter unwanted outliers.
3) Fix structural errors.
4) Fix missing data.
5) Validate the data.

This is a key step in ensuring compliance with AI legislation and regulation regarding IPR and copyrighted material along with ensuring no illegal material is also part of the training set.

#### 5.3.1.3 Data annotation and labelling

Data labelling and data annotation are similar but serve different purposes. Both terms are used interchangeably in some circumstances but are not the same process. Feeding a machine-learning model data is not enough for a computer to understand how to analyse and process it. Annotations and labelling describe data so that these algorithms can decipher it.

Annotations in machine learning are metadata used to describe the data. Machine learning uses large quantities of unstructured data to output meaningful information, and annotations provide every element of input information used by computing processes. For example, a picture with various elements uses annotations to define identifiable objects in the picture so that algorithms can understand and identify the same elements in future input.

Labelling is similar, but it is used to define data types. Input into an algorithm could be text or a picture, but a computing system does not know the difference between input types unless a user tell it. Data labelling tags both input types so that algorithms can decipher between the two and use them to establish patterns. In a picture, it tells algorithms what type of data is present such as a human or an animal. Labelling data is critical in Natural Language Processing (NLP) to help algorithms identify aspects of human communication, including words spoken, accents, and dialects.

With data annotation and labelling, there is a risk of exposing sensitive and/or personal identifiable information. There are different techniques to minimize the risk of exposing personal information. These include:

- Federated learning is the way to train AI models without centralizing data. Instead of collecting all the data in one place, models are given the ability to be trained directly on devices.
- Differential privacy is the way of processing data in a way that safeguards its confidentiality. When a developer is going to extract useful information from a group of data, they add a little random "noise" to each piece. Therefore, even if someone recognizes their data in this overall analysis, they will not be able to determine exactly what information belongs to other people.

## 5.4 Use of Open-Source Models

There is a risk with open-source models in which users feed their training material. So, if organizations limit harm from their models that are fully compliant with regulations, there will always be alternative sources for malicious actors to make use of GenAI. For example, the malicious actors may feed the model scraped data of a targeted company's e-mails in order to automate the production of large amounts of spam/phishing e-mails.

## 5.5 Purposeful Data Poisoning

There is a trend of purposeful data poisoning. This is generally done by people who make a living from creative works, who have started purposefully adding hidden features to the posted work. This is to protect it from being successfully used to train GenAI models [i.22]. For example, text-to-image generators work by being trained on large datasets that include millions or billions of images.

Some generators are only trained with images that the generator's maker owns or has a licence to use. But other generators have been trained by indiscriminately scraping online images, many of which may be under copyright. This has led to a slew of copyright infringement cases where artists have accused big tech companies of stealing and profiting from their work.

This is also where the idea of "poison" comes in. Researchers who want to empower individual artists have created tools to fight back against unauthorized image scraping. The tools work by subtly altering an image's pixels in a way that wreaks havoc on computer vision but leaves the image unaltered to a human's eyes. If an organization then scrapes one of these images to train a future AI model, its data pool becomes "poisoned". This can result in the algorithm mistakenly learning to classify an image as something a human would visually know to be untrue. As a result, the generator can start returning unpredictable and unintended results.

Developers hope that these tools will make big tech companies more respectful of copyright, but it is also possible that users could abuse the tool and intentionally upload "poisoned" images to generators to try and disrupt their services.

There are different approaches to mitigate against this problem.

Firstly, it is to pay greater attention to where input data are coming from and how they can be used. Doing so would result in less indiscriminate data harvesting. This includes only using data that has been licensed or has permission to use or own.

Secondly, to use technological fixes, for example, "ensemble modelling", where different models are trained on many different subsets of data and compared to locate specific outliers. This approach can be used not only for training but also to detect and discard suspected "poisoned" images.

Thirdly, to make use of audits. One audit approach involves developing a "test battery" - a small, highly curated, and well-labelled dataset – using "hold-out" data that are never used for training. This dataset can then be used to examine the model's accuracy.

# 6 Harmful Impacts from GenAI

## 6.1 Overview

Generative AI has a multitude of issues and risks pertaining to, but not limited to, the distribution of harmful content, copyright and legal exposure, sensitive information disclosure, amplification of existing bias, data provenance, lack of explicability/explainability and interpretability and hallucinations. These are harms which can be directed at targeted people, for example, deepfakes to exhort or indirectly, such as producing incorrect information which is then acted upon. It should be noted that all the clauses listed below can be carried out without the use of GenAI, but what GenAI has enabled is a large increase in the scale of these harms and a reduction in the time required to carry out these malicious actions. Also, when common languages are used, GenAI will do better at generating harmful content, as there is vastly more training material available than niche languages.

## 6.2 Prompt Injection Attack

### 6.2.1 Overview

A prompt injection is a type of cyberattack against Large Language Models (LLMs). The user (generally a malicious actor) disguises malicious inputs as legitimate prompts, manipulating generative AI systems (GenAI) into leaking sensitive data, spreading misinformation, or worse. Prompt injection can also be a way to test the model. The most basic prompt injections for example, make an AI chatbot ignore system guardrails and say things that it should not be able to. Prompt injections take advantage of a core feature of generative artificial intelligence systems, their ability to respond to users' natural-language instructions.

### 6.2.2 Direct Prompt Injection Attack

Direct prompt injections occur when the prompt is entered intentionally by the user where users attempt to manipulate the behaviour of Large Language Models directly through their User Input. Some techniques can be utilized for direct prompt injection. An example technique is jailbreaking, where the intended outcome is for the LLM to output malicious content which somehow bypasses their instructions and alignment training. Another technique is Prompt Leakage attempts. These can also have the intended effect of revealing the system prompt or the instructions to the interface of the LLM, which are meant to be hidden from the end user and dictate the model's behaviour.

### 6.2.3 Indirect Prompt Injection Attack

Indirect prompt injection is the insertion of malicious information into the data sources of a GenAI system by hiding instructions in the data it accesses, such as incoming emails or saved documents. Unlike direct prompt injection, it does not require direct access to the GenAI system, instead presenting a risk across the range of data sources that a GenAI system uses to provide context [i.23].

When a GenAI system gains access to emails, personal documents, organizational knowledge and other business applications, there is a marked increase in the scope to introduce malicious disinformation through indirect prompt injection using hidden instructions. A key component of hidden instructions comes from the fact that a GenAI assistant does not read data in the way that a human does. This makes it possible to devise exceedingly simple methods of insertion that are invisible to the human eye but are central to a GenAI system's retrieval process. When combined with the range of input methods available to a GenAI assistant, such as emails, documents and external web pages, the attack surface is broad and varied. This can create a risk for example, of external manipulation, exfiltration of data, phishing scams, injecting executable code or spreading disinformation.

Different measures can be implemented to mitigate the risk including maintaining good data hygiene, evaluating systems before deployment, providing user training and implementing technical guardrails.

## 6.3 Misinformation

Misinformation occurs when LLMs, used to produce GenAI, produce false or misleading information that appears credible. This vulnerability can lead to security breaches, reputational damage, and legal liability [i.34]. GenAI has accelerated the production of misinformation by making it faster, more scalable, and easier to create. Unlike human-generated misinformation, which requires time and effort to ideate and write, AI can generate misleading content in mere seconds. It can then be swiftly disseminated across multiple social media platforms. Often, the sheer volume of such content overwhelms traditional fact-checking systems, making it increasingly difficult for people to verify the accuracy of the information they see online [i.33].

One of the causes of misinformation is hallucination, when the LLM generates content that seems accurate but is fabricated. Hallucinations occur when LLMs fill gaps in their training data using statistical patterns, without truly understanding the content. As a result, the model may produce answers that sound correct but are completely unfounded. Also, biases introduced by the training data and incomplete information can contribute.

A related issue is overreliance. Overreliance occurs when users place excessive trust in LLM-generated content, failing to verify its accuracy. This overreliance exacerbates the impact of misinformation, as users may integrate incorrect data into critical decisions or processes without adequate scrutiny.

As well, human behaviour can cause misinformation to spread widely, even if the AI-generated content is identifiable. This happens because whether content is generated by AI or humans, users often share it without verifying its authenticity, which perpetuates the spread of misinformation. Users often do not distinguish between human- and AI-generated content. While AI-generated content may seem less credible, its structure, clarity, and flawless presentation can make it equally appealing to share.

Cognitive biases and social pressures can exacerbate this issue. Cognitive biases, such as confirmation bias, play a significant part in why individuals are more likely to believe and share information that aligns with their pre-existing beliefs, regardless of its veracity. Social media platforms, where people are often influenced by their peer networks, amplify this effect, making it easier for misinformation to spread rapidly. Moreover, the emotional tone of misinformation, whether it incites fear, anger, or urgency, tends to elicit stronger reactions, encouraging users to share without critical evaluation.

Examples of risk from misinformation:

1) Factual Inaccuracies - The model produces incorrect statements, leading users to make decisions based on false information.

2) Unsupported Claims - The model generates baseless assertions, which can be especially harmful in sensitive contexts such as healthcare or legal proceedings.

3) Misrepresentation of Expertise - The model gives the illusion of understanding complex topics, misleading users regarding its level of expertise.

4) Unsafe Code Generation - The model suggests insecure or non-existent code libraries, which can introduce vulnerabilities when integrated into software systems.

There are several types of strategies to reduce the risk of misinformation:

1) Retrieval-Augmented Generation (RAG) - Aims to enhance the reliability of model outputs by retrieving relevant and verified information from trusted external databases during response generation. This helps mitigate the risk of hallucinations and misinformation.

2) Model Fine-Tuning - Techniques such as Parameter-Efficient Tuning (PET) and chain-of-thought prompting can help reduce the incidence of misinformation.

3) Cross-Verification and Human Oversight - Encourage users to cross-check outputs with trusted external sources to ensure the accuracy of the information. Implement human oversight and fact-checking processes, especially for critical or sensitive information.

4) Automatic Validation Mechanisms - implementing methods to automatically validate key outputs, especially output from high-stakes environments.

5) Risk Communication - Identify the risks and possible harms associated with LLM-generated content, then clearly communicate these risks and limitations to users, including the potential for misinformation.

6) Secure Coding Practices - Aim to prevent the integration of vulnerabilities due to incorrect code suggestions.

7) User Interface Design - encourage responsible use, such as integrating content filters, clearly labelling AI-generated content and informing users on limitations of reliability and accuracy. Be specific about the intended field of use limitations.

8) Training and Education - Help users understand the limitations of GenAI LLMs and the importance of independent verification of generated content, and the need for critical thinking.

## 6.4 GenAI Hallucinations

AI hallucination is a phenomenon wherein a Large Language Model (LLM), often a generative AI tool, perceives patterns or objects that are non-existent or imperceptible to human observers, creating outputs that are nonsensical or altogether inaccurate [i.21].

Generally, if a user creates requests from a generative AI tool, they desire an output that appropriately addresses the prompt (that is, a correct answer to a question). However, sometimes AI algorithms produce outputs that are not based on training data, are incorrectly decoded by the transformer or do not follow any identifiable pattern. In other words, it "hallucinates" the response. These misinterpretations occur due to various factors, including overfitting, training data bias/inaccuracy and high model complexity.

AI hallucination can have significant consequences for real-world applications. For example, a healthcare AI model might incorrectly identify a benign skin lesion as malignant, leading to unnecessary medical interventions. One significant source of hallucination in machine learning algorithms is input bias. If an AI model is trained on a dataset comprising biased or unrepresentative data, it may hallucinate patterns or features that reflect these biases.

AI models can also be vulnerable to adversarial attacks, wherein bad actors manipulate the output of an AI model by subtly tweaking the input data. In image recognition tasks, for example, an adversarial attack might involve adding a small amount of specially crafted noise to an image, causing the AI to misclassify it. This can become a significant security concern, especially in sensitive areas such as cybersecurity and autonomous vehicle technologies. Techniques such as adversarial training, where the model is trained on a mixture of normal and adversarial examples, can mitigate security issues.

Different measures can be taken to prevent or minimize the occurrence of hallucinations:

- Use high-quality training data - what data is used in the training will reflect in the output.
- Define the purpose the GenAI model will serve as well as any limitations on the use of the model – this will help the system complete tasks and minimize irrelevant, "hallucinatory" results.
- Use data templates - these provide a predefined format, increasing the likelihood that an AI model will generate outputs that align with prescribed guidelines.
- Limit responses - defining boundaries for AI models using filtering tools and/or clear probabilistic thresholds can improve the overall consistency and accuracy of results.
- Test and refine the system continually.
- Rely on human oversight - human oversight ensures that, if the AI hallucinates, a human will be available to filter and correct it.

## 6.5 Loss of Confidentiality

The risk of loss of confidentiality from GenAI can occur because the Large Language Models (LLMs), the backbone of many GenAI systems, can inadvertently or maliciously leak sensitive information. This can occur through various means, such as data breaches, inadvertent disclosures, or sophisticated cyberattacks that exploit vulnerabilities within the AI systems.

Data leakage and privacy violations can happen as GenAI systems often require vast amounts of data to function effectively. This data, if not properly managed, can lead to significant privacy breaches. For instance, confidential business information or personally identifiable information (PII) might be exposed during AI training or inference processes. This can cause organizations to violate the regulatory landscape surrounding data privacy, such as GDPR and CCPA. Use of Shadow GenAI (unsanctioned or ad hoc generative AI use within an organization that's outside IT governance) also presents another avenue of risk where data leakage or compliance breaches can occur.

Data Protection Authorities such as Commission Nationale de l'Informatique et des Libertés (CNIL) have provided recommendations [i.35] to promote the responsible use of AI while ensuring compliance with personal data protection. These recommendations confirm that GDPR requirements are sufficiently balanced to address the specific challenges of AI. They provide concrete and proportionate solutions to inform individuals and facilitate the exercise of their rights:

- When personal data is used to train an AI model and may potentially be memorized by it, the individuals concerned need to be informed.

  The way this information is provided can be adapted based on the risks to individuals and operational constraints. Under the GDPR, in certain cases - especially when AI models rely on third-party data sources and the provider cannot contact individuals directly - organizations may limit themselves to general information (e.g. published on their website). When multiple sources are used, as is common with general purpose AI models, a broad disclosure indicating the categories of sources or listing a few key sources is generally sufficient.

- European regulations grant individuals the right to access, rectify, object and delete their personal data.

However, exercising these rights can be particularly challenging in the context of AI models, whether due to difficulties in identifying individuals within the model or modifying the model itself. AI developers should incorporate privacy protection from the design stage and pay special attention to personal data within training datasets by:

- Aim to anonymize models whenever it does not compromise their intended purpose.

- Incorporate solutions to prevent the disclosure of confidential personal data by AI models.

## 6.6 Malicious Code Generation

Current GenAI systems lack the specific capabilities and training necessary to independently create operational malware. Nonetheless, cybercriminals are using GenAI to uplift skills and refine existing malware, augment social engineering attacks, and provide 'malware-as-a-service'. Rather than enabling disruptive change, generative AI allows threat actors to move faster and at higher volume. For skilled actors, generative AI tools provide a helpful framework. For less skilled actors, they also provide a learning and productivity tool, enabling them to more quickly develop tools and incorporate existing techniques [i.36]. There are different applications in which GenAI can be used for malicious code generation [i.37]. These are:

- Techniques allowing malware to alter its code when it executes or rewrite itself entirely.

- GenAI agents potentially writing their own payloads or creating tools to overcome novel challenges.

- Agents could adapt tactics in real-time, allowing them to work remotely with less need for direction.

- Multiple agents working in cooperation, providing a persistence technique that allows constant learning and adaptation.

- Agents could reason about their environment and adapt their communications to 'blend in'.

As LLMs and GenAI models become more capable at producing computer code, malicious actors will use them to produce malicious code, as this follows the trend of bad actors being early adopters of new technologies. Though it is likely that a certain level of technical expertise will still be required to use AI as a constituent part of a successful cyber-attack [i.38].

NOTE: GenAI is not just used for malicious code generation. It has become the normal practice for coding software in general. This has largely replaced the use of script libraries.

To mitigate against the threat of malicious code generation, organizations should regularly control and evaluate whether their malware solutions are adapted to meet the threats posed by AI and other technological developments. This includes employee security awareness training to increase the knowledge of AI-enabled cyberattacks.

## 6.7 Spam Generation

### 6.7.1 Overview

Spam is any kind of unwanted, unsolicited digital communication that gets sent out in bulk. Often, spam is sent via email, but it can also be distributed via text messages, phone calls, or social media. With generative AI, malicious actors can now send phishing emails that bridge language barriers, reply in real time, and almost instantly automate mass personalized campaigns that make it easier to gain access to sensitive data.

### 6.7.2 Phishing

An AI phishing attack leverages artificial intelligence to make the phishing emails more convincing and personalized. A malicious actor could use AI algorithms to analyse vast amounts of data on a target segment, such as social media profiles, online behaviour, and publicly available information, which allows them to create personalized campaigns.

The phishing message could even include familiar touches, such as references to a user's recent purchases, interests, or interactions. This level of personalization increases the likelihood of success. AI can also easily generate convincing replicas of legitimate websites, making it difficult for the recipient to distinguish between the fake and real sites.

There are key principles that AI phishing is built on:

1) Data Analysis: The attacker uses algorithms and tools to scour the internet for vast amounts of data on the target group or individual. This includes social media profiles, public records, and online activities. They then analyse this data to understand the target's interests, behaviours, and preferences.

2) Personalization: With the collected data, AI generates highly personalized phishing emails. These emails may reference recent purchases, hobbies, or specific events in the target's life. This level of personalization makes the emails appear more legitimate and increases the likelihood of the victim falling for the scam.

3) Content Creation: Then, AI is used to generate convincing email content that mimics the writing style of the target's contacts or known institutions. This helps in creating a sense of familiarity and trust and overcomes any hurdles caused by language barriers.

4) Scale and Automation: Finally, AI makes it easy for attackers to scale their operations efficiently. They can generate numerous unique phishing emails in a short amount of time and use AI to target a wide range of individuals or organizations while also using AI to generate code, assist with triggering automations, and set up webhooks and integrations.

### 6.7.3 Mitigations

There are different best practices that businesses and users should take to prevent and detect not just AI spam but spam in general. These are not just steps that should be done once but need to be regularly refreshed and updated as the threat continues to evolve.

These include, but are not limited to:

- Conduct security awareness training. Cover traditional and new phishing attack techniques during security awareness training to ensure employees know how to identify phishing scams.

- Know the warning signs. Check for classic phishing scam errors, including typos, incorrect email addresses and other mistakes, as well as suspicious emails that create a sense of urgency or that could be from an impersonator.
- Do not click links or download attachments. Scrutinize links and downloads from all senders, even trusted sources. Do not copy or paste links into browsers.
- Do not share data. Question any message that asks for personal, business or financial data.
- Require Multi-Factor Authentication (MFA) and other password security best practices. Avoid sharing passwords and follow password hygiene guidelines.
- Use email security and anti-phishing tools. Email security gateways, email filters, antivirus and antimalware, firewalls, and web browser tools and extensions can catch many, but not all, phishing attempts. Use a layered security strategy.
- Adopt email security protocols. Email security protocols, such as SSL/TLS for HTTPS, SMTPS (SMTP Secure, essentially SMTP over TLS) and S/MIME (Secure/Multipurpose Internet Mail Extensions, a type of email encryption and authentication system), as well as email authentication protocols, including Sender Policy Framework (SPF), DomainKeys Identified Mail (DKIM) and Domain-based Message Authentication Reporting and Conformance (DMARC), can improve email security and help ensure email authenticity.

These methods are not foolproof, and when a phishing attack does succeed, there should be processes in place to minimize the harm and recover from it. Along with processes to learn and implement fixes to reduce the likelihood of it occurring again. Due to GenAI becoming better at mimicking the correct styles or masquerading as phishing, some of our traditional training to spot false emails and messages is no longer effective in the same but at the moment, the current training for spotting phishing is still an important first step in preventing it.

## 6.8 Deepfakes

### 6.8.1 Overview

A type of threat falling under the greater and more pervasive umbrella of synthetic media utilizes a form of artificial AI/ML to create believable, realistic videos, pictures, audio, and text of events which never happened [i.31].

Deepfakes are videos, audio, or images that seem real but have been manipulated with AI. They have been used to try to influence elections and to create non-consensual pornography, and also used to develop child sexual abuse material, low-cost deepfake adverts and synthetic terrorist content. The malicious use of deepfakes can cause an erosion of trust in elections, spread disinformation, undermine national security and empower harassers.

It can combine both sexual harassment and cyberbullying, which can be incredibly traumatic and humiliating for those who are targeted. Become objectified without consent and remove the ability to control who sees it, which could have serious consequences in real life: maybe their family sees it, maybe their boss sees it, maybe their significant other sees it, and maybe they all think it is real. Individuals could be threatened with sextortion or revenge porn generated by AI as a means of control and abuse, and many lack the financial or legal resources to seek justice in these situations.

**Table 1: Overview of different categories of risks associated with deepfakes**

| Psychological Harm | Financial Harm | Societal Harm |
|---|---|---|
| • (S)extortion<br>• Defamation<br>• Intimidation<br>• Bullying<br>• Undermining trust | • Extortion<br>• Identity theft<br>• Fraud (e.g. insurance/payment)<br>• Stock-price manipulation<br>• Brand damage<br>• Reputational damage | • News media manipulation<br>• Damage to economic stability<br>• Damage to the justice system<br>• Damage to the scientific system<br>• Erosion of trust<br>• Damage to democracy<br>• Manipulation of elections<br>• Damage to international relations<br>• Damage to national security |

### 6.8.2 Detection and Prevention

There are two distinct approaches to deepfake detection: manual and automatic detection. Manual detection requires a skilled person to inspect the video material and look for inconsistencies or cues that might indicate forgery. A manual approach can be feasible when dealing with low quantities of suspected materials, but it is not compatible with the scale at which audio-visual materials are used in modern society [i.32].

Automatic detection software can be based on a (combination of) detectable giveaways, some of which are AI-based themselves:

- Speaker recognition
- Voice liveness detection
- Facial recognition
- Facial feature analysis
- Temporal inconsistencies
- Visual artefacts
- Lack of authentic indicators

Several important cautions need to be kept in mind. One caution is that the performance of detection algorithms is often measured by benchmarking it against a common dataset with known deepfake videos. Also, detection evasion techniques using simple modifications in deepfake production techniques can drastically reduce the reliability of a detector.

Another problem detectors face is that audio-visual material is often compressed or reduced in size when shared on online platforms such as social media and chat apps. The reduction in the number of pixels and artefacts that sound, and image compression create can interfere with the ability to detect deepfakes.

While these will not stop the problem of deepfakes, they have the potential to reduce the harm and spread of deepfake material.

### 6.8.3 Reporting and Removal

Victims of deepfakes, especially non-consensual sexually explicit media attacks, often note the difficulty of removing content from all potential sources. Victims describe the recurring nightmare of having the same content appear on multiple sites and the frustration/difficulties of having to solicit formal actions every time before it can be removed. Sign posting and ease of finding the right resources or places to report and to have this content removed are valuable.

List of Potential Resources:

- The Federal Bureau of Investigation's (FBI) Internet Crime Complaint Centre (IC3) https://www.ic3.gov/. This is from the USA.
- Report inappropriate content and abuse on social media platforms using the platforms' reporting procedures, though this becomes more difficult if the targeted victim does not have an account for that platform.
- If a victim is under 18 years of age, incidents can be reported to the National Centre for Missing and Exploited Children via their cyber tip line at https://report.cybertip.org. This is based in the USA.
- Google's Help Centre, a resource available via Google that enables victims to remove fake pornography from Google searches https://support.google.com/websearch/answer/9116649?hl=en#:~:text=You%20or%20your%20authorized%20representative,representative%20submit%20in%20the%20form.
- StopNCII.org is a free tool designed to support victims of Non-Consensual Intimate Image (NCII) abuse and their partner network of victim advocates and non-profits around the world: https://stopncii.org/partners/global-network-of-partners/

# 7 GenAI Content and Material

## 7.1 How GenAI is shared and spreads online

As with any user-created content, GenAI content and material are shared on all types of social media and related platforms. This difference is mainly the scale and amount of content that can be shared. For example, a digital artist uploads a finished piece once a week. A person using GenAI could upload hundreds of pieces of AI art every day. This distorts the recommendation algorithms that power online services, leading to GenAI content pushing out non-GenAI content. This becomes a more serious problem when bots are used to spread misinformation and misleading content.

The bigger problem is that many of the harmful deepfakes, especially non-consensual images, come from open-source systems or systems built by state actors, and they are disseminated on end-to-end encryption messaging platforms, where they are far harder to trace and remove. This means content is shared between people without an easily identifiable source. It often comes to light when a person is caught doing something else, and any material is discovered on their devices being shared on encrypted messaging platforms.

The sharing of GenAI content online is not necessarily a harmful or illegal act. It is the intent and context of the GenAI content and material being shared online which determines if it is harmful or illegal.

## 7.2 Best Practice Measures within GenAI Platforms / Services

### 7.2.1 Prevention by Design

Prevention involves efforts to limit the creation of harmful deepfakes. This can include adopting prompt filters to prevent models from being instructed to create certain types of content (e.g. nude content); removing harmful content from model training datasets; and blocking outputs before they are presented to users [i.42].

Prevention measures consist of any attempt to stop a harmful deepfake from being created and typically involve introducing safeguards 'upstream' to limit what models can produce. Preventative measures can include:

1) Training datasets: Model developers could opt to omit certain types of data from their training datasets. For example, a firm developing an image model could seek to identify and remove "Not Safe For Work" (NSFW) images from training datasets, which could make it more difficult for their technology to generate sexual deepfakes. Similarly, model developers could remove content from their datasets that depicts public figures and celebrities, thereby making it more challenging for users to create deepfakes that portray such individuals.

2) Prompt filters: Model developers could introduce filters that instruct a model to reject problematic prompt requests. For example, a prompt filter could be set up to reject an instruction to 'create a nude version of this photo' or variations of that kind. The number of terms included in a list of prohibited prompts can vary depending on the specific model and its intended use.

3) Output filters: In addition to adding filters at the front end of a model where the user inserts prompts, model developers could choose to add output filters that automatically inspect generated content and block that which is deemed harmful. For example, a firm developing a text-to-image model could use AI-powered image classifiers to identify and block nude content, which might be used in sexual deepfake imagery.

Preventative measures can be an efficient way to stop deepfake content from being created in the first place, rather than expending resources in identifying that content once it has begun to circulate online. These measures are particularly well-suited to tackling the creation of sexually explicit deepfakes (i.e. deepfakes that demean), as well as some types of defrauding deepfakes where a user's intent to defraud is clear. However, preventative measures have limitations, which include:

- It can be challenging for model developers and other actors upstream to know when a user intends to create harmful content. For example, a user may seek to create an image or video of a celebrity, public figure, or politician for purely satirical purposes. Likewise, a user may want to ask questions of a model that relate to hate and terror for educational reasons. Prompt and output filters can struggle to distinguish between benign requests like these and requests that carry malicious intent. In some cases, it is impossible to know whether a given piece of content amounts to a deepfake until it is shared with others, and even then, that judgment call is not straightforward.

- Preventive measures may be less effective when applied to open-source models. This is because third-party actors can usually modify these models, including by removing preventative measures installed by the original model developer. Third-party actors could also further train (or 'finetune') a model with harmful content, meaning that the model is more likely to create similar content in future.

- Preventative measures are not always robust. Even the most well-crafted preventative interventions have weak spots. While filters can stop many attempts to create harmful content, bad actors often still find ways to circumvent them. For example, in the case of the Taylor Swift deepfake incident, while Microsoft Designer had prompt filters to prevent the generation of content featuring public figures, users were able to generate sexually explicit content featuring the singer by misspelling her name in prompts. Similarly, although red teaming can be a valuable way of stress testing models, even the most extensive exercises will not be able to identify every vulnerability in a model, not least because there are a wide variety of 'jailbreaking' techniques that bad actors could deploy.

### 7.2.2 Metadata

Metadata (data about data) provides descriptive information about a piece of content, for example, about its author, creation or modification date, and the tools used to create or modify it. Unlike watermarking, which involves marking the content itself, metadata is instead added to a file that accompanies the content. Many organizations, for example, have now signed up to the Coalition for Content Provenance and Authenticity (C2PA) scheme, which has been described as a 'nutritional' label for content [i.41]. C2PA is a specification that "addresses the prevalence of misleading information online through the development of technical standards for certifying the source and history (or provenance) of media content." To this end, verifiable information may be cryptographically embedded into images, videos, audio, and some types of documents in a way that is difficult to remove and that makes tampering evident. More broadly, the International Organization for Standardization (ISO) is finalizing the International Standard Content Code (ISCC), a universal content identifier that transparently fingerprints content across platforms using hashing.

C2PA and ISCC have the potential to improve our capacity for identifying deepfakes and distinguishing real from fake content. A deepfake video that has been embedded with a watermark at the point of its creation stands a greater chance of being identified as synthetic than one without. Equally, a genuine audio file that has been embedded with provenance metadata is more likely to have its authenticity verified than one that lacks the same inscription.

However, there are limitations to Metadata that include, but are not limited to:

- C2PA and ISCC will not be implemented by every model developer or deployer. As noted, some model developers and deployers are deliberately designing tools to create harmful content (e.g. as in the case of nudify apps). It is extremely unlikely that these actors will voluntarily choose to label their content or attach metadata or watermarks.

- Bad actors will attempt to remove C2PA and ISCC information. Those intent on causing harm to others, be it by circulating deepfake adverts or deepfake political content, will always look to eliminate any signal that the content is not genuine. Some metadata fields, for example, can be easily manipulated or removed by bad actors who use file editing or metadata editing tools.

- C2PA and ISCC may be less effective for addressing deepfakes that demean. In the case of sexualised deepfakes or those that contain content intended to bully a victim, supplying contextual information to a viewer may help to prove that the deepfake is false; however, the harmful impact of the deepfake can remain the same.

- Watermarks can be unintentionally weakened through editing. Content creation and sharing is a messy and convoluted process, often involving multiple rounds of editing, downloading, compression and sharing. Although embedding techniques are becoming more robust, content alterations like these can still damage watermarks and make them harder to detect.

- Widespread adoption of C2PA and ISCC could result in genuine content being called into question. As techniques like labels and metadata become more popular, users of online platforms may expect to see these signals on content as standard and may raise questions where they are not visible. This could lead to perverse outcomes, including cases where authentic content is viewed as fake because it lacks metadata to demonstrate its provenance. Indeed, it may allow individuals who are genuinely depicted in an unflattering circumstance to dishonestly claim that the content in question is a deepfake (a phenomenon known as the 'liar's dividend').

### 7.2.3 Red Teaming

Red teaming is a type of offensive security evaluation approach [i.40] that, in the context of AI, seeks to find vulnerabilities in AI models. Put simply, this involves 'attacking' a model to see if it can generate harmful content. The red team can then seek to fix those vulnerabilities by introducing new and additional safeguards, for example, filters that can block such content.

Red team exercises involve inputting a series of prompts to a model to see whether it generates harmful content. Red teaming is a bespoke and tailored activity, with prompts varying from model to model and from exercise to exercise. Although every exercise differs, red teaming tends to be dynamic in the sense that evaluators can adjust their prompts depending on the results that are coming up (e.g. to lean into probing for one type of harm if it appears to be a vulnerability from initial prompting).

This means red teaming a couple is useful for flexibility, meaning it can be scaled up and down to suit the given context. Along with adaptability, it can be adjusted to changing user behaviours and emerging risks.

It should also be noted that red teaming has several limitations, which include:

- Red teaming is more difficult for video, audio, and multi-modal models. Audio-visual and multimodal models produce a greater volume and variety of content for every input prompt, which tends to make outputs more difficult to analyse. For example, to red team a video model would often require both a visual and audio assessment of the outputs.
- Human error can lead to inaccurate assessments of model outputs. Human reviewers, particularly those with minimal experience, may miss or misjudge harmful content produced by a model during red team assessments. While some red teamers use automated classifiers to support the review of model outputs, these too are liable to inaccurately assess content.
- Red teaming does not fully replicate real-world uses of a model. Red teaming is often conducted within a controlled environment, which means that evaluations do not always mirror real-world applications once the model has been released.
- The results of red teaming exercises are not easily compared. Unlike benchmark tests, where the same prompts are entered into every model, red team exercises are designed to be customized, with different attacks used for different models. While this has its advantages, it also makes it difficult to compare the results of one assessment with another.

## 7.3 Tackling the Content Shared from GenAI Platforms

### 7.3.1 Detection

Detection encompasses efforts to distinguish real from fake content, even where no contextual data has been attached to that content. This means using tools to reveal the origins of content, regardless of whether information has been attached or embedded with watermarks or metadata. These efforts are primarily undertaken by online platforms and can involve the use of both automated and human-led content reviews. Detection methods include the use of forensics, hash matching and user reporting [i.42].

Forensic techniques involve the use of machine learning systems or human review to recognize telltale signs that content is wholly or partially synthetic. These techniques vary by content type. For example, to determine whether images are synthetic content, reviewers could look for a lack of symmetry in facial attributes, as well as erroneous lighting or shadows. For videos, content reviewers can look at whether an individual featured in the content is blinking and moving their head naturally or unnaturally. Identifying fake audio content is more challenging, but there are still signals that can be monitored, for instance, by looking for inconsistencies in waveforms that could suggest alterations or tampering. While humans can perform many of these forensic techniques, they cannot always do so at the speed and scale that online platforms require. To address this challenge, there are detection tools that promise to automate this process.

Hashing is an umbrella term for techniques that create a 'fingerprint' of a given piece of content. In practice, this means using an algorithm to analyse content and create a 'hash' that can represent it. Hashes are then stored in a database that can be accessed by multiple parties as required. In the context of online safety, online platforms can use hashing to notify other platforms of illegal or harmful content they have identified, and vice versa. Hashing databases exist for CSAM, terror content, and non-consensual intimate images. Similar databases could, in theory, be created for known deepfake content, such as political deepfakes, where that is not already being captured by existing hashing schemes.

Detection methods have their limitations, including:

- Bad actors will always seek to adjust their methods to outmanoeuvre forensic techniques.
- Identifying deepfakes requires more than just knowing whether content is synthetic. It also involves a determination of whether that content is intended to misrepresent or cause harm.
- Content editing can diminish the accuracy of deepfake detection tools.
- Hashing techniques can be vulnerable to 'collision'. This describes circumstances where different content has the same hash value, which could result in genuine content being identified as a deepfake (or vice versa).
- Users of online platforms can find it challenging to identify deepfake content.

### 7.3.2 Enforcement

Enforcement involves setting clear rules about the types of synthetic content that can be created and shared on online services. It also involves acting against users who breach those rules, for example, by suspending or removing user accounts [i.42].

Online platforms and GenAI service providers can take enforcement action where their rules are breached. This includes:

- Issuing warnings to users - Users can be issued a warning or a 'strike', notifying them that they have breached the rules. Some platforms offer policy training to their users after an initial warning.
- Taking down content - For example, where the content clearly violates a platform's terms of service.
- Suspending or removing users - User accounts can be suspended or docked. Some model developers choose to move offending users onto a restricted version of their service that has more limited capabilities. User accounts can also be terminated entirely.
- Labelling content where there is not a clear breach. Where there is not a clear breach, online platforms may make a decision to label content or models. In the case of Hugging Face, for content that is not fully prohibited, the platform can request model owners to add a 'Not for All Audiences' tag to their models, or to 'gate' the model to make it less visible to others.

Establishing and enforcing clear rules makes it less likely that users will be able to create and share deepfake content. Clear terms of service, community guidelines and licence agreements can reduce the ability of bad actors to exploit loopholes, whilst also enabling content moderators to make fairer and more informed decisions as they review content. However, there are limitations to enforcement:

- Policies can suffer from arbitrary boundaries. The rules set by some online platforms may appear to be incoherent.
- Policies can lack specificity. The terms of service and community guidelines of some firms in the technology supply chain can be too generic, for example, prohibiting the creation or sharing of 'harmful content' without providing detailed examples of what that means in practice.
- Licence agreements are difficult to enforce in the case of open-source models. Once an open-source model has been released, it is difficult to monitor who is using it and for what purposes. Even where a model developer is aware of their models being misused to create prohibited deepfake content, they may be able to step in and block that behaviour.

### 7.3.4 Reporting

In addition to relying on tools to proactively detect deepfakes, online platforms can also invite their users to report this content, which their content review teams can then review and take action on as necessary. Some online platforms already enable their users to report content which could include illegal or harmful deepfake content. Under regulations such as the EU AI Act [i.6], there are additional requirements for companies to comply with, which include reporting to an authorized regarding misuse of their AI services and products.

Depending on the country where a platform operates and the user's location, there may be additional requirements to report to the country's law enforcement agencies or police. This should occur if a victim is under 18 years of age, and their dedicated organizations, child protection organizations, should be reported to as well.

### 7.3.5 Removal

The main method of removal of nonconsensual GenAI content is through the use of hashes to identify and remove that content. Currently, this only applies to images, but research is underway to extend it to video and audio as well. In the case of video and audio, a report has to be made to the platform where that content is hosted, and they would take it down, which is a manual process, while for images, it can be done automatically. Also, requesting that web links to that material be removed can aid in limiting the spread of the material by making it harder to find. If required, companies may have to record and retain data when removing such material if served by warrants. Requirements for this can be found in ETSI TS 102 232 [i.43] Series.

A non-exhaustive list of resources to aid in the removal of non-consensual deep fake content:

- https://takeitdown.ncmec.org/
- https://stopncii.org/
- https://www.ic3.gov/
- https://revengepornhelpline.org.uk/
- https://www.inhope.org/EN#hotlineReferral

# 8 Conclusion

## 8.1 Overview

There are patterns that link the threats from GenAI, the mitigations and the compliance to legislation and regulation. For developers of GenAI, by understanding the measures they have to comply with they can implement controls and polices. Often these controls and polices also mitigate harm from GenAI by making it harder for them to be used to commit harm. See Figure 1 below.

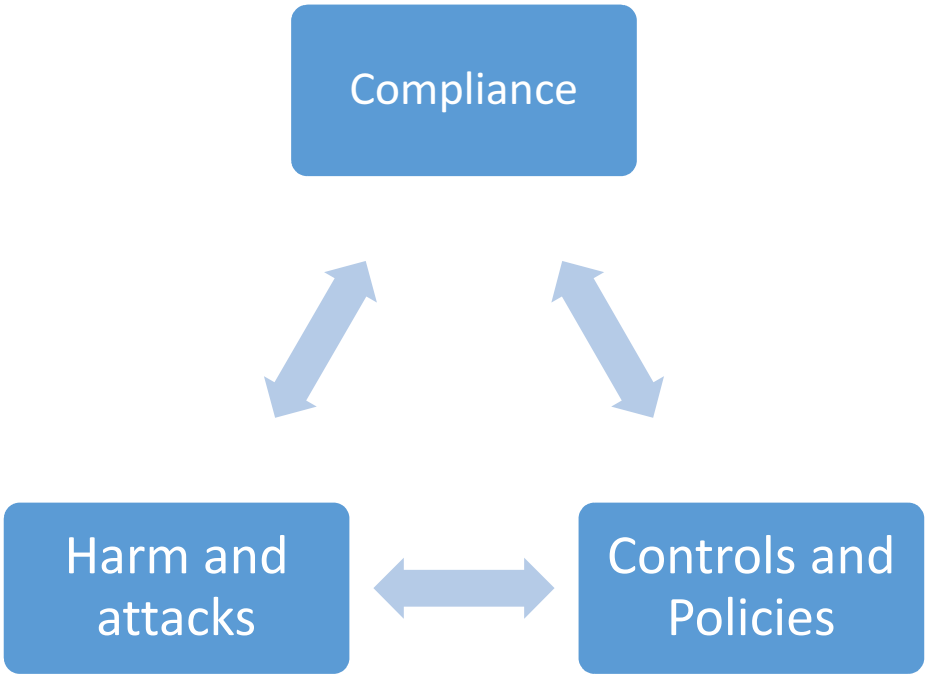


**Figure 1: Connections**

An example of this many countries that have passed laws and regulations related to AI have requirements to prevent the spread of misinformation and have transparency of the data sources used to train the GenAI. This requires developers to know what data they are using. So, they should not blindly scrape data from anywhere. If they use copyrighted material, they need permission. Also, if the training material has personal identifiable information, it should be anonymized.

Along with prompt filters and controls to prevent users from misusing the GenAI to create harmful content. These steps are not exhaustive but are key to minimizing the risk of GenAI from being used to leak personal information, which leads to a loss of confidentiality and prevents the spread of misinformation, which can be harmful by ensuring the information being presented by the GenAI can be verified.

## 8.2 Trustworthy AI

### 8.2.1 Overview

Trustworthy AI is built upon three pillars that form the foundation of trustworthy AI as indicated in Figure 2, and necessitate adherence throughout the entire AI system lifecycle as presented in ETSI TS 104 119 [i.39]:

- *"Lawful: AI systems shall rigorously comply with all applicable legal and regulatory frameworks. This encompasses adherence to national, international, and European Union legislation, including but not limited to the General Data Protection Regulation (GDPR) [i.48] and relevant sector-specific directives. This adherence ensures AI operations remain within established legal parameters, safeguarding fundamental rights and societal values."*
- *"Ethical: Beyond strict legality, AI systems are required to embody and uphold established ethical principles and values. This component is instantiated through four core ethical principles:*
  - *Respect for Human Autonomy: AI systems should augment human capabilities, facilitate informed decision-making, and preserve human control.*
  - *Prevention of Harm: AI systems shall be designed to preclude the infliction of physical, psychological, or economic detriment. Proactive identification and mitigation of potential negative impacts are imperative.*
  - *Fairness: AI systems shall operate equitably, actively mitigating unjustifiable bias and discrimination, thereby ensuring impartial treatment across individuals and groups.*

- *Explicability: The processes, functionalities, and decision-making mechanisms of AI systems shall exhibit transparency, interpretability, and comprehensibility to relevant stakeholders, thereby enabling scrutiny and accountability."*

- *"Robust: AI systems are required to possess both technical and societal robustness. This necessitates that they be reliable, secure, and resilient, capable of consistent and safe operation within diverse real-world environments, while also adapting responsibly to evolving societal contexts. Technical robustness pertains to attributes such as accuracy, dependability, and cybersecurity, whereas societal robustness encompasses broader ethical considerations and societal impact."*

The Trustworthy AI pillars can be mapped to GenAI to help mitigate the potential for harm associated with it.

NOTE: These same principles for GenAI being trustworthy can be applied similarly to Agentic AI.

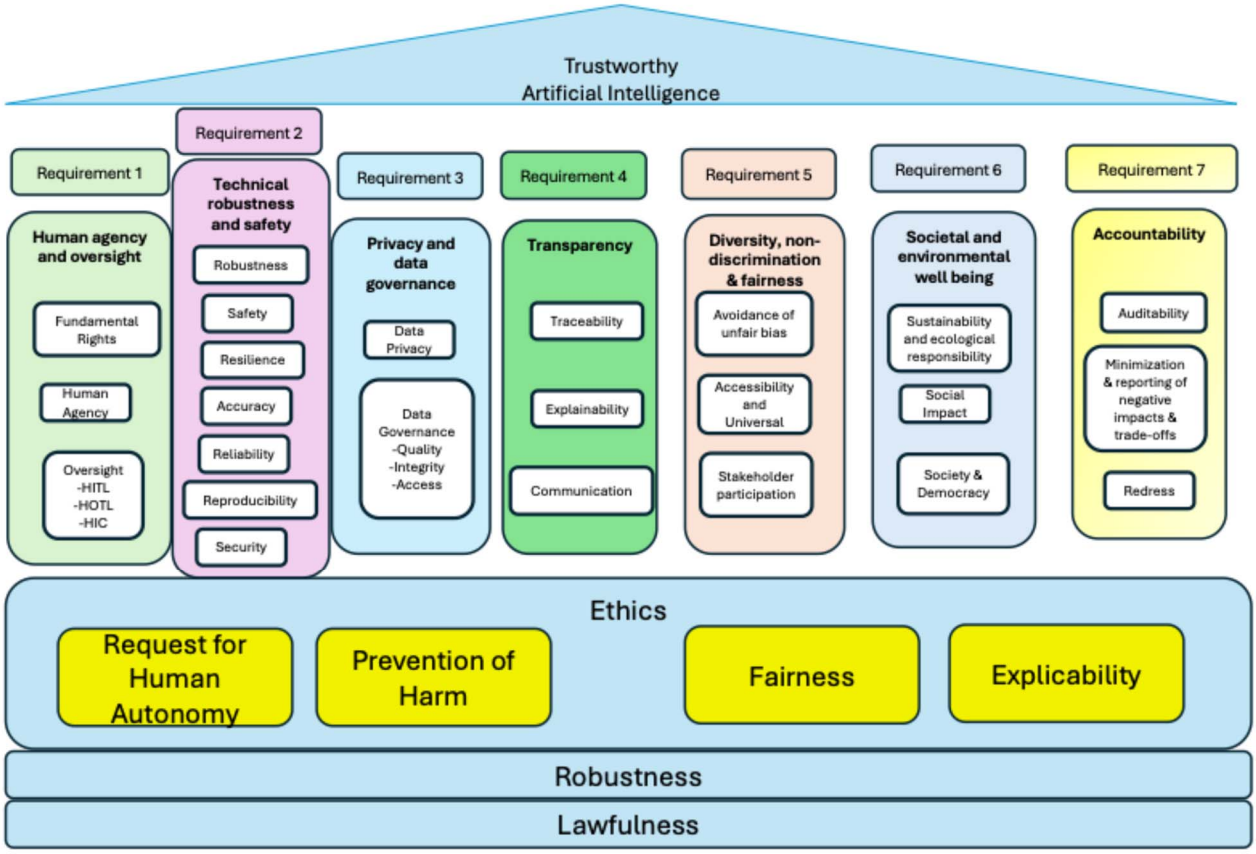


**Figure 2: Trustworthy AI pillars, requirements and characteristics**

## 8.2.2 Mapping of GenAI to Trustworthy AI

### 8.2.2.1 Overview

The requirements of trustworthy AI apply to GenAI. By showing that a GenAI system meets the requirements for trustworthy AI, it can aid in meeting many of the principles and requirements under the various regulations and legalisations shown in clause 4.3.

### 8.2.2.2 Table of Mapping

The controls and methods for mapping of GenAI to Trustworthy AI are drawn from clauses 4 to 8 inclusive.

**Table 2: Mapping of GenAI to Trustworthy AI**

| | **Requirement 1 Human Agency and oversight** | **Requirement 2 Technical robustness and safety** | **Requirement 3 Privacy and data governance** | **Requirement 4 Transparency** | **Requirement 5 Diversity, non-discrimination & fairness** | **Requirement 6 Societal and environmental well being** | **Requirement 7 Accountability** |
|---|---|---|---|---|---|---|---|
| **GenAI controls and methods for requirements** | Have a human oversight process in place to be able to respond to incorrect outputs and hallucinations. | Test and refine the system continually. | Use of federated learning | Data used to train the algorithm has been legally accessed or licensed. | Training and Education to help users understand the limitations of GenAI LLMs. | User Interface Design should encourage responsible use, such as integrating content filters, clearly labelling AI-generated content and informing users on limitations of reliability and accuracy. | Risk Communication. |
| | Cross-Verification with trusted external sources to ensure accuracy of outputs. | Model fine tuning. | Use of differential privacy methods. | From ETSI TS 104 119 [i.39]: "*The processes, functionalities, and decision-making mechanisms of AI systems shall exhibit transparency, interpretability, and comprehensibility to relevant stakeholders, thereby enabling scrutiny and accountability*". | Ensuring diversity/inclusion for example adopting universal design, ensuring accessibility, and providing relevant stakeholders with education and support. | From ETSI TS 104 119 [i.39]: "*AI systems shall rigorously comply with all applicable legal and regulatory frameworks. This adherence ensures AI operations remain within established legal parameters, safeguarding fundamental rights and societal values*". | Organizations understand their role with regard to the systems they develop or manage, put in place appropriate risk management systems, and share information with other organizations as needed. |

| | Requirement 1 Human Agency and oversight | Requirement 2 Technical robustness and safety | Requirement 3 Privacy and data governance | Requirement 4 Transparency | Requirement 5 Diversity, non-discrimination & fairness | Requirement 6 Societal and environmental well being | Requirement 7 Accountability |
|---|---|---|---|---|---|---|---|
| | Human Oversight and Monitoring - System use is monitored after deployment, and updates are implemented as needed to address any risks that materialize. | Use of Retrieval-Augmented Generation. | Aim to anonymize models whenever it does not compromise their intended purpose. | | | Respect for Human Autonomy: AI systems should augment human capabilities, facilitate informed decision-making, and preserve human control. | |
| | | Use of Automatic Validation Mechanisms. | Incorporate solutions to prevent the disclosure of confidential personal data by AI models. | | | | |
| | | Secure Coding Practises | | | | | |
| | | Continuous research to mitigate societal, safety and security risks and investment in effective mitigation measures when identified. | | | | | |
| | | Systems are subject to risk assessments, and mitigations needed to ensure safe operation are put in place before deployment. | | | | | |

# Annex A:
# Change history

| Date | Version | Information about changes |
|---|---|---|
| 03 2025 | 0.0.1 | Skeleton draft |
| 06 2025 | 0.0.2 | Early draft |
| 09 2025 | 0.0.3 | Stable draft |
| 09 2025 | 0.0.4 | Stable draft |
| 10 2025 | 0.0.5 | Stable draft |
| 11 2025 | 0.0.6 | Final draft |

# History

| Version | Date | Status |
|---|---|---|
| V1.1.1 | January 2026 | Publication |
| | | |
| | | |
| | | |
| | | |